**信息**就是把那些我们不清楚的事情阐明描述，而已经明确或者只晓的东西让我们再“知晓”一遍，这些被知会的内容就不再是信息了。

**数据挖掘**---首先是有一定量的数据作为研究对象，挖掘--顾名思义，说明有一些东西不是放在表面上一眼就能看明白，要进行深入的研究、对比、甄别等工作，最终从中找到规律或知识。

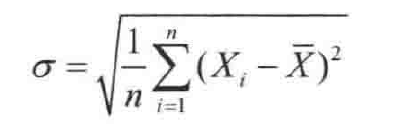
**机器学习**---先想想人类学习的目的是什么，是掌握知识、掌握能力、掌握技巧，最终能够进行比较复杂或者高要求的工作。类比一下机器，我们让机器学习，不管学习什么，最终目的都是让它独立或者至少半独立地进行相对复杂或者高要求的工作。更多的是让机器帮助人类做一些大规模的数据识别、分拣、规矩总结等人类做起来比价吃力的事情，但是和人工智能 相差还是甚远。

概率论和线性代数与其相关性很大；每门学科的发展最终能够降低人类认知或描述世界的成本，带来工作效益的直接或间接的提升.有了这个思路,我们在理解很多现象的时候都会更加自然.

概率是对大量样本分布比例的解释,而不是猜测.

------------------------------30页

**标准差**，公式：



**加权均值**：平均值这种指标有一个‘兄弟’-加权均值。权指的是权重，也就是指所占的“比重”或者“重要程度”。

**众数**：在样本对象中出现最多的那个数字。

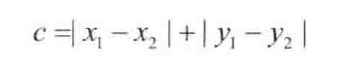
**中位数**：顾名思义，就是位于中间位置的数字

**欧式距离**：例如 每门课的成绩减去平均分，再把差值平方。

For num in numList:

(Num – average(numList))\*\*2

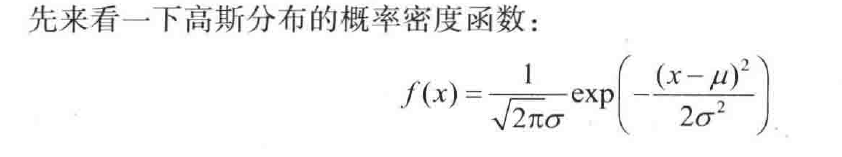
**曼哈顿距离**：

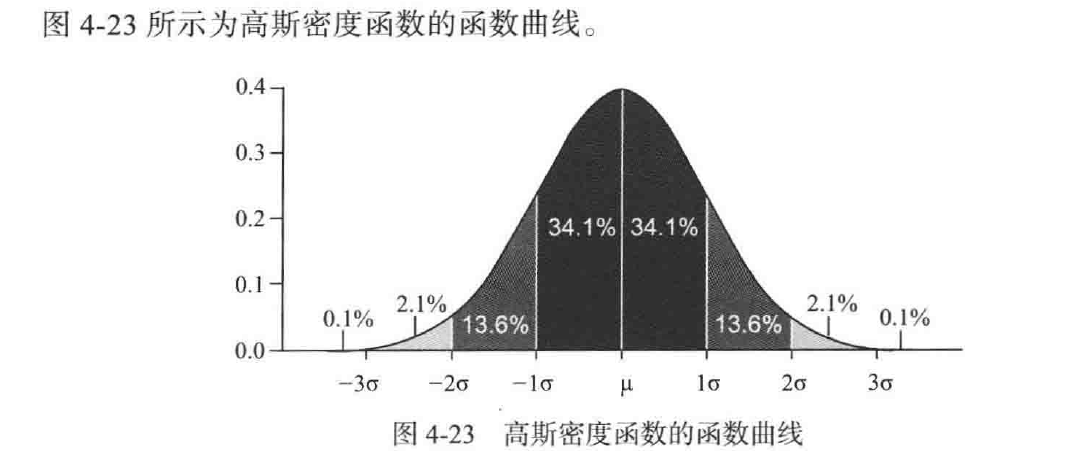


可用于地图上距离的检测

**抽样**：是一种非常好的了解大量样本空间分布情况的方法，样本越大则抽样带来收益越明显。

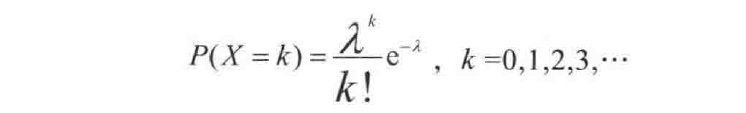
**正态分布**，又名**高斯分布**，它作为分布特性的一种，首先是用来描述统计对象的，如果统计对象的分布特性符合高斯分布，那么所有针对高斯分布的定理和“经验值”都能够直接套用。特点：“一般般的很多，极端的很少“。



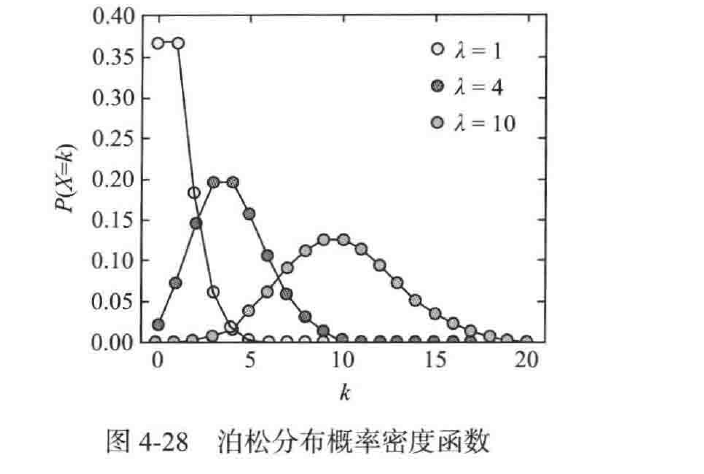


# exp指以自然常数e为底的指数函数

**泊松分布**是一种统计与概率学中常见的离散概率分布。它适合与描述单位时间内随机事件发生的次数。

# 是单位时间（或单位面积）内随机事件的平均发生率。

# k! 指k的阶乘

----------50